Finding an Optimal Inversion Median: Experimental Results

Adam Siepel and Bernard M.E. Moret

Department of Computer Science
University of New Mexico
Albuquerque, NM 87131, USA

acs@ncgr.org and moret@cs.unm.edu

Overview

- whole-genome evolution
- genomic distances and computational characteristics
- why is the median of three genomes interesting?
- breakpoint median vs. median under general evolutionary distances
- a branch-and-bound approach to median finding
- *experimental results*: how do the two medians differ?
- *experimental results*: uniqueness of the inversion median
- *experimental results*: speed and parallel implementation
- newer work: GRAPPA with exact medians and distance corrections

Whole-Genome Evolution

- Gene content and gene order carry significant phylogenetic signal.
- Gene order is affected by *rearrangement* events.
- Rearrangement events include
  - *inversion*: pick up a subsequence of genes and put it back in place, but reversed
  - *transposition*: pick up a subsequence of genes and insert it somewhere else in the genome
  - *inverted transposition*: inversion combined with transposition
- Rearrangement events are relatively rare, so they give information about deep evolutionary history.

Genomic Distances for Single Genomes

- *Breakpoint distance*: the number of adjacencies present in one genome but not in the other

  not an evolutionary distance, but model-independent

- *Inversion distance*: the number of inversions

  plausible for many organellar genomes (esp. chloroplasts)

- *ITT (Inversion, Transposition, and Inverted Transposition) distance*: requires weights for the 3 events

  fits the generalized Nadeau-Taylor model

Current Results for Genomic Distances

- *Breakpoint distance*: trivially computable in linear time
- *Inversion distance*: computable in linear time for signed genomes (Moret et al.), but NP-hard for unsigned genomes (Caprara)
- *Transposition distance*: 1.375-approximation (Hannenhalli and Pevzner), complexity unknown
- *ITT distance*: complexity unknown, approximations for various special cases (Eriksen, others)

Edit distances are used for the evolutionary distances. Correcting edit distances to obtain expected evolutionary distances is generally advantageous (Moret et al., Wang and Warnow, Wang)

The Median Problem

Given three genomes A, B, and C, construct a fourth genome M (the median) such that the distance

$$d(A, M) + d(B, M) + d(C, M)$$

is minimized.

Used in phylogeny reconstruction to produce tentative ancestral genomes.

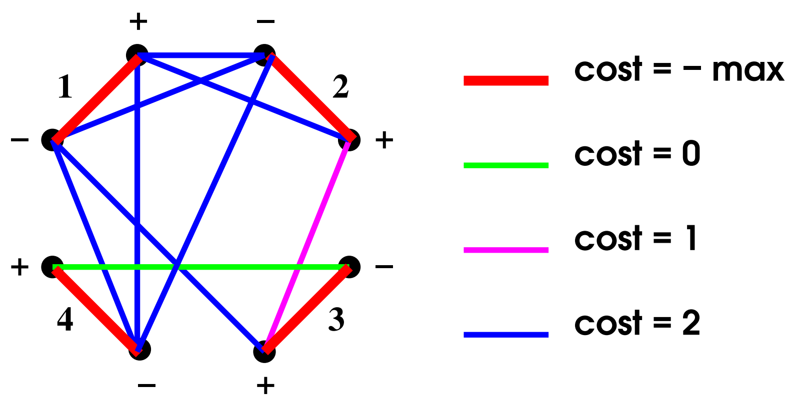
NP-hard for both breakpoint and inversion distances.

Median for Breakpoints: A Simple TSP

Blanchette and Sankoff gave a reduction of the Median Problem for Breakpoints to the Travelling Salesperson problem (TSP).

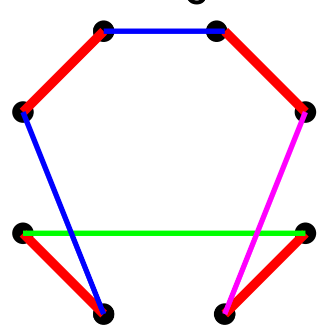
Genes become cities; the distance between two cities i and j is $(3-k)$, where $0 \leq k \leq 3$ is the number of genomes that include the subsequence $i\ j$ or $-j\ -i$.

+1 −2 +4 +3
+1 +2 −3 −4
+2 −3 −4 −1



edges not shown have cost = 3

Adjacency A B becomes an edge from A to −B



an optimal solution corresponding to genome +1 +2 −3 −4

Median for Inversions: A Very Different Beast

No practical formulation of the optimization problem is known.

Good heuristics with guaranteed error bounds are not known.

Only general methods are used: branch-and-bound, integer-linear programming, etc.

Tight upper bounds are missing, so the search can be expensive.

A Branch-and-Bound Framework

Theorem: The median score $S(\Pi)$ of a set of three permutations $\Pi = \{\pi_1,\ \pi_2, \pi_3\}$, separated by pairwise distances $d_{1,2}, d_{1,3}$, and $d_{2,3}$, obeys these bounds:

$$\left\lceil \frac{d_{1,2} + d_{1,3} + d_{2,3}}{2} \right\rceil \le S(\Pi) \le \min\left\{ (d_{1,2}+d_{2,3}), (d_{1,2}+d_{1,3}), (d_{2,3}+d_{1,3}) \right\}$$

The upper bound uses one of the three permutations as the median.

The lower bound is a special case of the old "twice around the tree" bound for the TSP and uses only the triangle inequality.

Corollary: Let π_1, π_2, π_3 and ϕ be permutations with ϕ on a shortest path between π_1 and a median M. The median score $S(M)$ obeys

$$\left\lceil \frac{d_{2,\phi} + d_{3,\phi} + d_{2,3}}{2} \right\rceil \le S(M) - d_{1,\phi} \le \min\{(d_{2,\phi}+d_{2,3}), (d_{2,\phi}+d_{3,\phi}), (d_{2,3}+d_{3,\phi})\}$$

Why the Framework Uses Only Metric Properties

- the distance function $d(\, , \,)$ in the theorem can be any metric, so the result applies to all genomic distances, not just inversion
- the bounds are very easy to compute if one has a way to compute pairwise distances efficiently
- the bounds are fairly tight for small evolutionary distances

The Branch-and-Bound Algorithm

- uses the two theorems for bounding and for branch selection
- searches alternatives of equal cost in depth-first search manner (to get an early first solution)
- as specialized for inversion: generates neighbor candidates by considering all $\binom{n}{2}$ possible inversions
- uses a hash table (based on hashing the permutations) to store the frontier nodes

Experiments: Goals

- measure number of permutations generated in a search
- compare inversion scores of breakpoint median and inversion median
- compare numbers of distinct breakpoint medians and of distinct inversion medians

Experiments: Setup

- implemented in C using the linear-time inversion distance computation of `GRAPPA`

- generated triples of genomes parameterized by number of genes n and number of inversions per tree edge i

- a triple is generated from the identity permutation (the "true" median) by applying i randomly selected (uniform over all $\binom{n}{2}$ choices) inversions on each edge

Parallelizing the Code

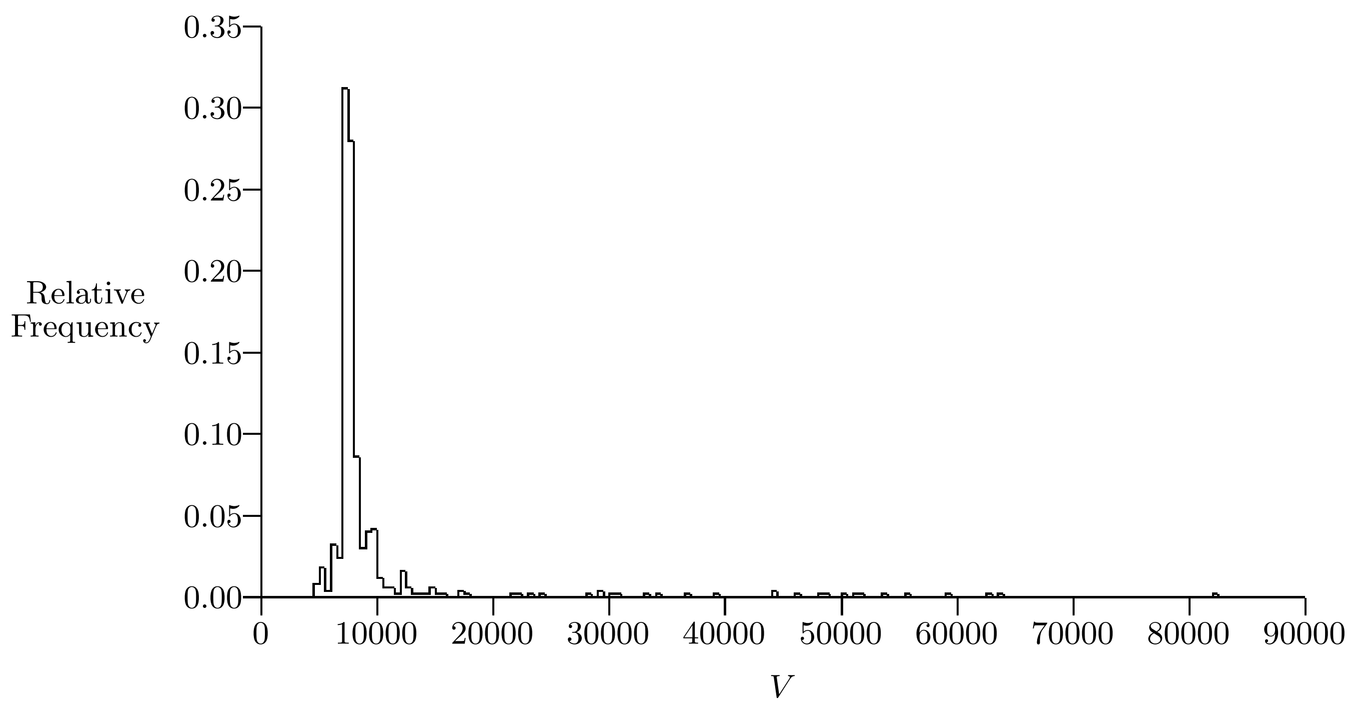
- A parallel implementation may enable us to solve real-world instances within realistic time.
- The code is easy to parallelize in terms of assigning work to threads.
- Branch-and-bound requires shared access (read and write) to the hash table, the priority queue, and the storage pool.
- The priority queue has lock bypassing (Zhang, ACM JEA), while the hash table has fine-grained readers-writers locks on a bucket-by-bucket basis. The storage pool has a simple global lock.

Performance of Bounds

In 87% of the experiments, fewer than 10,000 vertices were visited and in 95%, fewer than 20,000 were visited.

The number of visited vertices grows quadratically with the number of genes and linearly with the number of inversions.

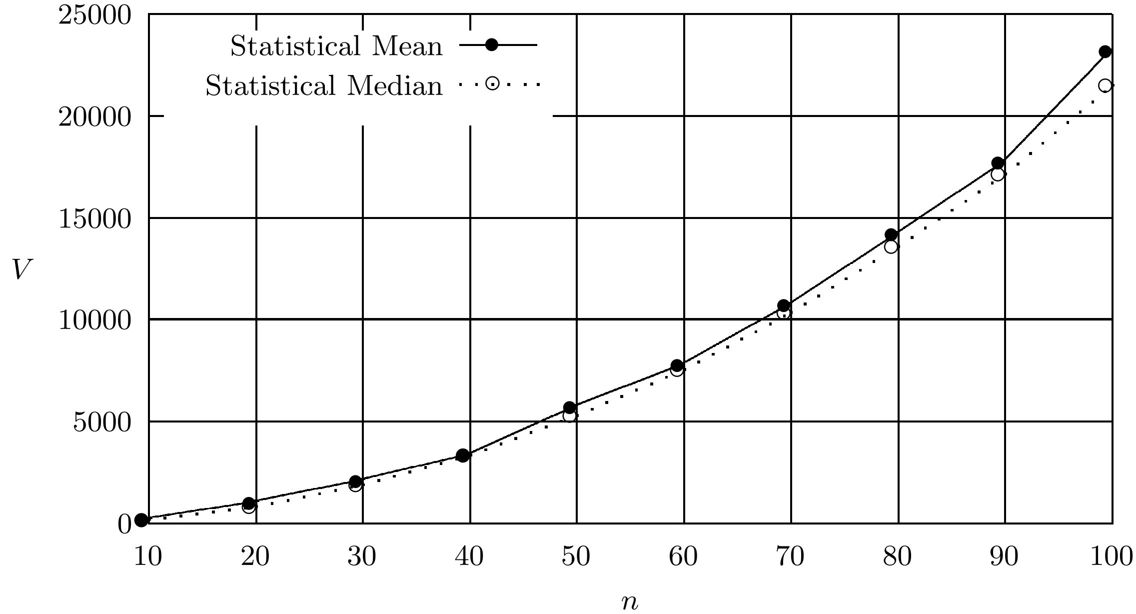
Number of vertices visited for $n = 50$ *and* $i = 7$*:*



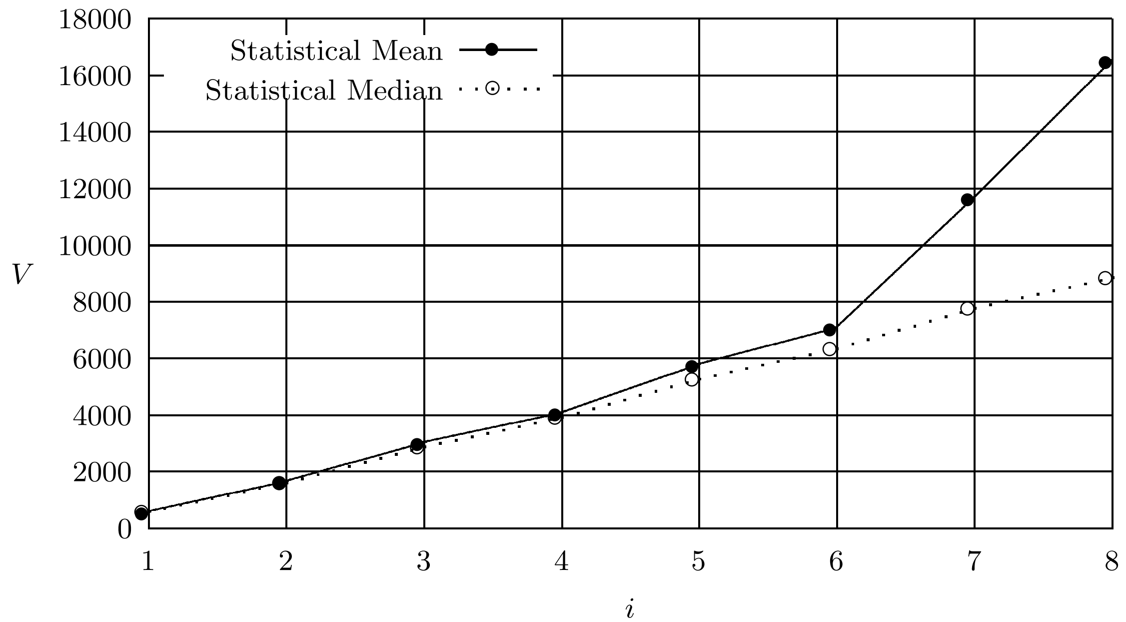
Performance of Bounds (cont'd)

Average number V of vertices visited (50 datasets)

for $i = 5$ *and* $10 \leq n \leq 100$*:*



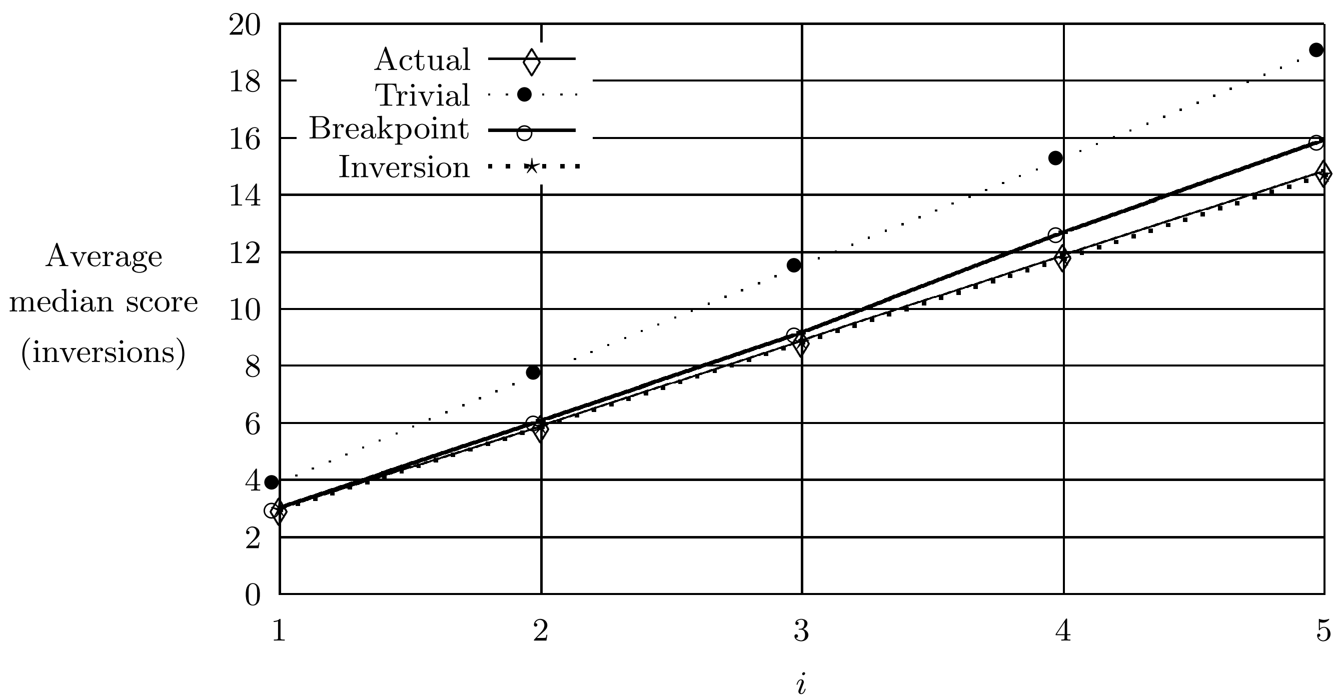
for $n = 50$ *and* $1 \leq i \leq 8$*:*



Comparing Breakpoint Median and Inversion Median

Inversion scores of inversion medians were 5–15% lower than those of breakpoint medians.

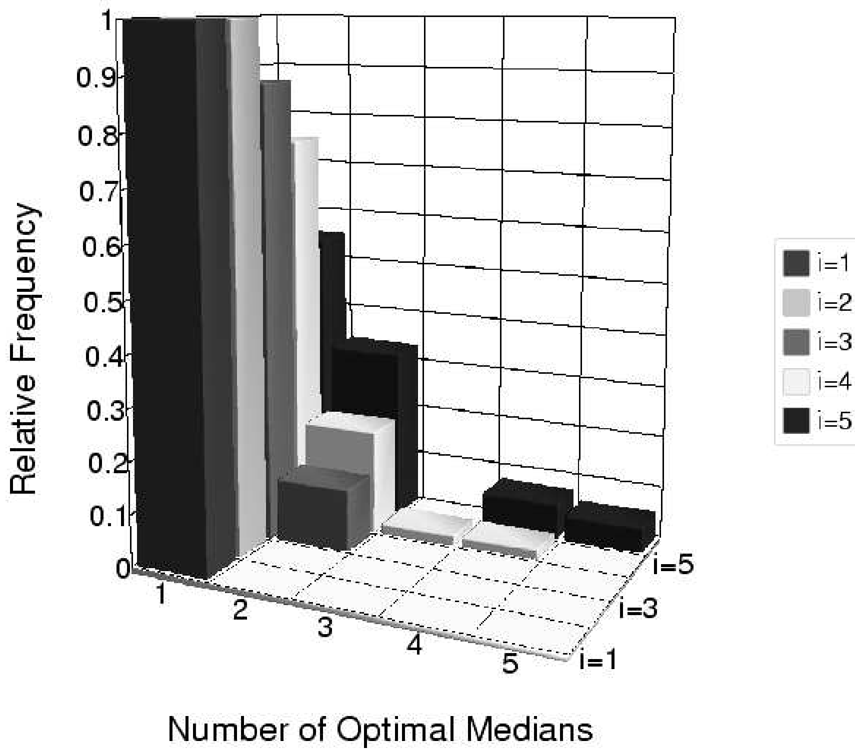
25 *genes*, 50 *runs:*



Uniqueness of Inversion Medians

The breakpoint median is almost never unique, while the inversion median is very often unique.

Number of optimal medians for $n = 15$, $1 \leq i \leq 5$:



Unique inversion medians tend to be "true" medians, whereas breakpoint medians tend to be "trivial" medians (one of the three original permutations).

Advantages in Phylogeny Reconstruction

We incorporated the inversion median code within `GRAPPA`.

- uniqueness of medians leads to higher confidence in optimal tree
- rarity of "trivial" medians reduces number of zero-length edges
- inversion scores of trees are lowered and thus better bounding becomes possible
- trees exhibit better topological accuracy

Conclusions

- general branch-and-bound framework for computing exact median under any distance metric
- implementation for inversion runs fairly fast for a wide range of realistic instances
- experiments show that inversion median is preferable to breakpoint median, both in absolute terms and within the context of phylogeny reconstruction
- current implementation can be refined in many ways and will be included in next major release of `GRAPPA`
- challenge is to devise efficient algorithms for computing other genomic distances (transposition, ITT, etc.)